Universidade Federal de Uberlândia
Faculdade de Matemática

Bacharelado em Estatística

REGRESSÃO LINEAR MÚLTIPLA NA MODELAGEM DE RESULTADOS NA *NATIONAL BASKETBALL ASSOCIATION* (NBA)

Luiz Felipe Vieira Maciel

Uberlândia-MG
2019

Luiz Felipe Vieira Maciel

REGRESSÃO LINEAR MÚLTIPLA NA MODELAGEM DE RESULTADOS NA *NATIONAL BASKETBALL ASSOCIATION* (NBA)

Trabalho de conclusão de curso apresentado à Coordenação do Curso de Bacharelado em Estatística como requisito parcial para obtenção do grau de Bacharel em Estatística.

Orientadora: Profa. Dra. Maria Imaculada de Sousa Silva

Uberlândia-MG

2019

Universidade Federal de Uberlândia
Faculdade de Matemática

Coordenação do Curso de Bacharelado em Estatística

A banca examinadora, conforme abaixo assinado, certifica a adequação deste trabalho de conclusão de curso para obtenção do grau de Bacharel em Estatística.

Uberlândia, ____ de _____ de 20____

BANCA EXAMINADORA

_____

Profa. Dra. Maria Imaculada de Sousa Silva

_____

Prof. Dr. Ednaldo Carvalho Guimarães

_____

Prof. Dra. Nádia Giaretta Biase

Uberlândia-MG

2019

Agradecimentos

Agradeço primeiramente a minha mãe Marluce e ao meu pai Mário, por todo amor, apoio, educação, incentivo, ensinamentos e dedicação. Todos esses atributos foram imprescindíveis para a minha trajetória até aqui.

Agradeço a minha madrinha Luci, que desde muito novo me incentivou a estudar e sempre buscar conhecimento.

A todos da minha família, aos avôs, avós, tios, tias, primos e primas agradeço por fazerem parte da minha vida. Agradeço também aos professores do ensino fundamental, médio e superior que a seu tempo, e cada um, à sua maneira, também contribuíram com a minha formação.

A todos os meus amigos e colegas que compartilharam esses anos de graduação comigo, agradeço pelo companheirismo e amizade.

"Ninguém ignora tudo. Ninguém sabe tudo. Todos nós sabemos alguma coisa. Todos nós ignoramos alguma coisa. Por isso aprendemos sempre."

Paulo Freire

Resumo

Os esportes em geral desempenham um papel muito importante na sociedade. Seus benefícios são percebidos tanto nas questões relacionadas com a saúde física e mental, quanto no entretenimento, além de muito importante economicamente para uma região. Diante deste fato, o uso de técnicas estatísticas adequadas, tanto na parte descritiva como na tomada de decisões torna-se uma ferramenta importante para o planejamento de ações que possibilitem a obtenção de melhores resultados. Neste contexto, este trabalho avaliou, por meio da estatística descritiva das informações coletadas e do método de análise de regressão linear múltipla, quais as variáveis regressoras, estatísticas do jogo de basquete, foram significativas para explicação da variável resposta, quantidade total de vitórias dos times da liga profissional de basquete dos Estados Unidos, a *National Basketball Association* (NBA). O modelo final apresentou um bom ajuste, conseguindo explicar 94% da variabilidade encontrada no número total de vitorias. Para a realização das análises, os dados foram coletados no *site* oficial da NBA ao final de 9 temporadas regulares (2009-2019). Todas as análises e resultados foram executados e obtidos nos softwares Excel e R.

Palavras-chave: Basquete; Regressão Linear Múltipla; Número de Vitórias; Distância de Cook.

Abstract

Sports in general play a very important role in society. Its benefits are realized both in terms of physical and mental health, as well as entertainment, and very important economically for a region. Given this fact, the use of appropriate statistical techniques, both descriptive and decision-making becomes an important tool for planning actions that enable better results. In this context, this work evaluated, through the descriptive statistics of the collected information and the method of multiple linear regression analysis, which regressive variables, statistics of the basketball game, were significant to explain the response variable, total number of team wins. of the professional basketball league in the United States, the National Basketball Association (NBA). Final model Showing a good fit, being able to explain 94 % of the variability found in the total number of wins. To perform the analyzes, data were collected on the official NBA website at the end of 9 regular seasons (2009-2019). All analyzes and results were performed and obtained using Excel and R.

Keywords: Basketball; Multiple Linear Regression; Numbers of Wins; Cook Distance.

Sumário

Lista de Figuras

Lista de Tabelas

1. Introdução

O basquete é a segunda modalidade mais popular do mundo [14], com uma abrangência muito grande em quase todos os países dos continentes ao redor mundo. Não é de se surpreender que quando se pensa nesse esporte, lembra-se quase que instantaneamente dos Estados Unidos, país onde originou-se o esporte [2]. Em consequência desse fato, faz-se referência à *National Basketball Association* (NBA), a liga profissional de basquete daquele país. Considerando todas as competições desportivas no mundo, essa liga possui um dos maiores faturamentos, ficando atrás apenas de três outras competições, sendo elas a *National Fotball League* (NFL), *Major League Baseball* (MLB) (ambas dos Estados Unidos) e a *Premier League* (campeonato inglês de futebol) [7]. Tais competições fazem com que todo o negócio seja muito lucrativo para os proprietários de equipes, para as cidades das equipes ou que sediam os jogos, além das várias emissoras, até de outros países, que transmitem os jogos, não deixando de mencionar que as competições são uma boa fonte de entretenimento para os moradores e turistas de todas as partes do mundo. O basquete é um jogo extremamente dinâmico e bastante complexo, com situações que variam a cada período do jogo. Logicamente nem todos os fatores podem ser mensurados em números, mas algumas destas características podem ser medidas através das várias estatísticas que descrevem o panorama de cada equipe, em cada jogo.

Durante a partida de basquete, cada equipe tem 5 jogadores, sendo que todos atacam e defendem, podendo serem substituídos quantas vezes for desejado. O tempo total de jogo é de 48 minutos, que são divididos em 4 períodos. Havendo empate, deve-se jogar uma prorrogação de 5 minutos, e se permanecer a igualdade da pontuação será necessário mais um tempo extra e assim sucessivamente até que haja um vencedor. Em cada ataque a equipe pode marcar 1, 2 ou 3 pontos.

Com o início no ano de 1949, após a unificação com outras ligas concorrentes, a competição da NBA foi se tornando cada vez mais popular com o tempo. A NBA atualmente, possui 30 franquias (29 estadunidenses e 1 canadense) sediadas em 28 cidades diferentes. Para a disputa do campeonato, os times são divididos em duas conferências (Oeste e Leste), contendo 15 equipes cada uma, de forma que cada equipe faça 82 jogos durante a temporada regular. Após todas estas partidas, os 8 melhores de cada conferência se classificam para os *playoffs*, jogando uma série de melhor de 7 jogos, ou seja, quem vencer 4 partidas avança para a próxima rodada. A grande final é um confronto entre o campeão da conferência do oeste e o campeão da conferência do leste.

A estatística é uma ferramenta de suma importância para análise dos resultados de uma com-

petição, sendo a modelagem por meio da técnica de modelo regressão linear múltipla (MRLM) um método eficiente para avaliar as covariáveis que de fato são significativas para explicar a variável resposta.

Para o caso da NBA, uma análise de regressão múltipla poderá ter um impacto significativo para avaliação de desempenho das equipes, podendo as conclusões das análises serem bastante valorizadas pelos comandantes das equipes, uma vez que estas informações podem ser utilizadas a seu favor para obtenção da vitória.

A técnica dos modelos de regressão já tem sido usada em outros trabalhos, inclusive utilizando dados da NBA. Em um destes trabalhos, utilizou-se uma ponderação das estatísticas do jogo em uma única equação, para a obtenção de uma taxa denominada de *Player Efficiency Rating* (PER) calculada a partir de daos de jogos da NBA. Um modelo de regressão linear foi utilizado para verificar a influência dessa taxa, e consequentemente das covariáveis que a compõem, para explicação da variável resposta,quantidade total de vitórias da equipe durante a temporada, obtendo-se um bom ajuste do modelo de regressão linear [16].

Nos dias atuais os desportos ganharam um novo formato. São parte de uma indústria em pleno desenvolvimento: a indústria do entretenimento esportivo. Como parte dessa indústria e seu principal produto é que o esporte assumiu um novo aspecto aos olhos de todos: de atividade e prática massificada, o esporte tornou-se um negócio dotado de um grande mercado e de um elevado potencial de venda e comercialização. Sendo visto como produto, o esporte ganha uma nova dimensão diante de seus analistas e gestores, tornando-se parte de uma complexa estrutura de mercado com suas características e peculiaridades [6].

Dado o crescimento dessa indústria esportiva, a análise estatística das variáveis envolvidas nos resultados dos jogos ou na estruturação das equipes, torna-se uma estratégia importante na tomada de decisões, visando sempre melhores e mais lucrativos resultados. A análise de regressão é uma técnica muito difundida para modelagem não apenas no basquete, como em outras modalidades. Como exemplo, podemos citar o futebol, que é o esporte mais popular do Brasil e para o qual, também pode ser aplicada a regressão linear múltipla para estimação de êxito nos resultados dos jogos [8].

Para o presente trabalho, será apresentada a técnica de modelagem de regressão linear múltipla aplicada às estatísticas de 9 temporadas regulares da NBA (2009-2019). O objetivo é encontrar as variáveis que descrevem de maneira significativa a quantidade de vitórias de uma equipe durante o campeonato. Os dados para realização do trabalho foram coletados no *site* oficial da NBA [8]. Realizando-se as tabulações, rotinas, análises e interpretações através do *software* R [10].

2. Fundamentação Teórica

2.1 Introdução ao Método de Regressão

Análise de regressão é uma técnica estatística para investigar e realizar uma modelagem entre as variáveis. As aplicações da regressão são inúmeras e ocorrem em quase todos os campos da ciência, incluindo a engenharia, a física, a química, na economia, administração, ciências biológicas, sociologia e nos esportes. De fato, a análise de regressão pode ser a técnica estatística mais amplamente utilizada [5].

Com esta técnica, é interessante conhecer os efeitos que algumas variáveis exercem ou não sobre outras. Mesmo que não exista explicitamente uma relação linear entre as variáveis, podemos relaciona-las por meio de uma expressão matemática, que pode ser útil para se estimar o valor de uma das variáveis quando conhecemos os valores das outras, sob determinadas condições [11].

Genericamente, tais relações funcionais podem ser representadas pela Equação 2.1:

$$Y = f(X_1, X_2, ..., X_k) \tag{2.1}$$

onde Y representa a variável resposta (dependente) e os X_h $(h = 1, 2, ..., k)$ são as variáveis regressoras (covariáveis).

São exemplos de relações entre variáveis que podem ser analisadas utilizando o método da regressão:

i) A pressão sanguínea dos pacientes (Y) de um hospital em função da idade (X);

ii) Relação entre as variação salarial (Y) ao decorrer de um período de tempo (X);

ii) A variação do preço de um produto (Y) conforme a percentagem de desconto (X_1), a quantidade ofertada (X_2), o custo sobre o produto(X_3).

2.2 Modelo de Regressão Linear Simples

De acordo com [11], dados n pares de valores de duas variáveis, $X_i, Y_i (i = 1, 2, ..., n)$, se admitirmos que Y é função linear de X, podemos estabelecer uma regressão linear simples, cujo modelo estatístico é representado pela Equação 2.2:

$$Y_i = \alpha + \beta X_i + \varepsilon_i \tag{2.2}$$

onde α e β são os parâmetros e ε_i o erro aleatório.

Ao estabelecer um modelo de regressão linear simples (MRLS), pressupomos que:

i) a relação entre X eY é linear;

ii) os valores de X são fixos, isto é, X não é uma variável aleatória;

iii) a média do erro é nula, isto é, $E(\varepsilon_i) = 0$;

iv) para um dado valor de X_i, variância do erro ε é sempre σ^2, denominada variância residual, isto é,

$$E(\varepsilon_i^2) = \sigma^2$$

ou

$$E[Y_i - E(Y_i|X_i)]^2 = \sigma^2$$

v) o erro de uma observação é não correlacionado com o erro em uma outra observação, isto é, $E(\varepsilon_i\varepsilon_j) = 0$ para $i \neq j$;

vi) os erros têm distribuição normal com média 0 e variância σ^2.

Considerando-se duas variáveis X e Y, com n pares $((X_1,Y_1,)(X_2, Y_2)$ $...(X_n, Y_n))$, se Y é função linear de X, pode-se estabelecer uma regressão linear. Tal relação pode ser observada na Figura 2.1:

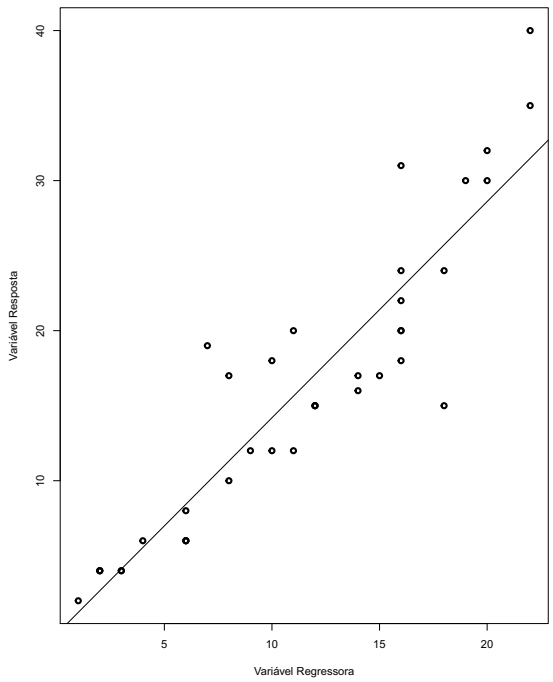


Figura 2.1: Gráfico de Regressão Linear Simples

Na Figura 2.1, pode-se visualizar uma representação de um MRLS. Percebe-se que a relação entre as duas variáveis pode ser descrita por uma reta, em que à medida que a variável regressora aumenta, a variável resposta acompanha o seu crescimento, ou seja, há uma relação diretamente proporcional. Pode-se haver também uma relação inversamente proporcional, ou até mesmo uma regressão linear perfeita, em que a reta passa por todos os pontos observados.

2.2.1 MÉTODO DE MÍNIMOS QUADRADOS

A etapa inicial da análise de regressão é, a obtenção das estimativas a e b dos parâmetros α e β da regressão. Os valores dessas estimativas podem ser obtidos a partir de uma amostra de n de valores $X_i, Y_i (i = 1, 2, ..., n)$, que correspondem a n pontos num gráfico, a partir dos quais, de acordo com [11], define-se a equação estimada do modelo a seguir:

$$\hat{Y}_i = a + bX$$

em que $\hat{Y}_i$, a e b são, respectivamente, as estimativas de $E(Y_i) = \hat{\alpha} + \hat{\beta}X_i$, α e β.

Para cada par de valores X_i, Y_i pode-se estabelecer o desvio

$$\epsilon_i = Y_i - \hat{Y}_i = Y_i - (a + bX_i)$$

O método dos mínimos quadrados consiste em adotar como estimativas dos parâmetros os valores que minimizam a soma dos quadrados dos desvios

$$Z = \sum_{i=1}^{n} \epsilon_i^2 = \sum_{i=1}^{n} [Y_i - (a + bX_i)]^2$$

A função Z terá mínimo quadrado quando suas derivadas parciais em relação a a e b forem nulas. Tais derivadas são representadas pelas Equações 2.3 e 2.4:

$$\frac{\partial Z}{\partial a} = -2 \sum_{i=1}^{n} [Y_i - (a + bX_i)] = 0 \tag{2.3}$$

$$\frac{\partial Z}{\partial b} = 2 \sum_{i=1}^{n} [Y_i - (a + bX_i)](-X_i) = 0 \tag{2.4}$$

Simplificando, tem-se sistema de equações normais:

$$\begin{cases} na + b \sum_{i=1}^{n} X_i = \sum_{i=1}^{n} Y_i \\ a \sum_{i=1}^{n} X_i + b \sum_{i=1}^{n} X_i^2 = \sum_{i=1}^{n} X_i Y_i \end{cases} \tag{2.5}$$

Resolvendo o sistema de Equações 2.5, tem-se :

$$a = \frac{\left(\sum_{i=1}^{n} X_i^2\right)\left(\sum_{i=1}^{n} Y_i\right) - \left(\sum_{i=1}^{n} X_i\right)\left(\sum_{i=1}^{n} X_i Y_i\right)}{n\sum_{i=1}^{n} X_i^2 - \left(\sum_{i=1}^{n} X_i\right)^2}$$

$$b = \frac{n\sum_{i=1}^{n} X_i Y_i - \left(\sum_{i=1}^{n} X_i\right)\left(\sum_{i=1}^{n} Y_i\right)}{n\sum_{i=1}^{n} X_i^2 - \left(\sum_{i=1}^{n} X_i\right)^2}$$

Para verificar que a fórmula para o cálculo de b pode ser escrita de diversos modos, quais sejam:

$$b = \frac{n\sum_{i=1}^{n} X_i Y_i - \left(\sum_{i=1}^{n} X_i\right)\left(\sum_{i=1}^{n} Y_i\right)}{n\sum_{i=1}^{n} X_i^2 - \left(\sum_{i=1}^{n} X_i\right)^2} = \frac{\sum_{i=1}^{n} X_i Y_i - \frac{\left(\sum_{i=1}^{n} X_i\right)\left(\sum_{i=1}^{n} Y_i\right)}{n}}{\sum_{i=1}^{n} X_i^2 - \frac{\left(\sum_{i=1}^{n} X_i\right)^2}{n}} =$$

$$\frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})Y_i}{\sum_{i=1}^{n}(X_i - \bar{X})^2} =$$

$$\frac{\sum_{i=1}^{n}(X_i(Y_i - \bar{Y}))}{\sum_{i=1}^{n}(X_i - \bar{X})^2} = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2} = \frac{\sum_{i=1}^{n} x_i Y_i}{\sum_{i=1}^{n} x_i^2} = \frac{\sum_{i=1}^{n} X_i y_i}{\sum x^2}$$

sendo

$$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n}, \bar{Y} = \frac{\sum_{i=1}^{n} Y_i}{n}, x_i = X_i - \bar{X}, y_i = Y_i - \bar{Y}$$

De acordo com [11], assinalemos duas relações importantes úteis que podem ser obtidas a partir das Equações 2.3 e 2.4:

Lembrando que $Y_i = (a + bX_i) = Y_i - \bar{Y} = \epsilon_i$, tais equações ficam:

$$\sum_{i=1}^{n} \epsilon_i = 0 \tag{2.6}$$

e

$$\sum_{i=1}^{n} X_i \epsilon_i = 0 \tag{2.7}$$

Temos, também, que

$$\sum_{i=1}^{n} \hat{Y}_i \epsilon_i = \sum_{i=1}^{n} (a + bX_i)\epsilon_i = a \sum_{i=1}^{n} e_i + b \sum_{i=1}^{n} X_i \epsilon_i$$

De acordo com as Equações 2.6 e 2.7, concluímos que:

$$\sum_{i=1}^{n} \hat{Y}_i \epsilon_i = 0 \tag{2.8}$$

As relações 2.6, 2.7 e 2.8 mostram, respectivamente, que:

i) a soma dos desvios é iguais a zero,

ii) a soma dos produtos dos desvios pelos correspondentes valores estimados da variável independente é igual a zero, e

iii) a soma dos produtos dos desvios pelos respectivos valores estimados da variável dependente é igual a zero.

2.3 Modelo de Regressão Linear Múltipla

Segundo [11], tem-se um modelo de regressão linear múltipla (MRLM) quando se admite que o valor da variável dependente é função linear de duas ou mais variáveis independentes. O modelo estatístico de uma regressão linear múltipla com k variáveis independentes é:

$$Y_j = \alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + ... + \beta_k X_{ki} + \epsilon_i, i = 1, ..., n$$

ou

$$Y_j = \alpha + \sum_{i=1}^{k} \beta_i X_{ij} + \epsilon_j \tag{2.9}$$

Para determinar os estimadores que minimizam a soma dos quadrados dos resíduos utiliza-se o Método dos Mínimos Quadrados, recomendado devido à sua precisão. O MRLM com k variáveis independentes, pode ser descrito na forma matricial de acordo com a Equação 2.10:

$$y = X\beta + \varepsilon \tag{2.10}$$

em que:

$$Y = \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ . \\ . \\ . \\ y_n \end{bmatrix}, \quad X = \begin{bmatrix} 1 & x_{11} & x_{21} & . & . & . & x_{k1} \\ 1 & x_{12} & x_{22} & . & . & . & x_{k2} \\ 1 & x_{13} & x_{23} & . & . & . & x_{k3} \\ . & . & . & . & & & . \\ . & . & . & & . & & . \\ . & . & . & & & . & . \\ 1 & x_{1n} & x_{2n} & . & . & . & x_{kn} \end{bmatrix}, \quad \beta = \begin{bmatrix} \beta_0 \\ \beta_2 \\ \beta_3 \\ . \\ . \\ . \\ \beta_k \end{bmatrix} \quad e \quad \varepsilon = \begin{bmatrix} \varepsilon_0 \\ \varepsilon_2 \\ \varepsilon_3 \\ . \\ . \\ . \\ \varepsilon_n \end{bmatrix}$$

De acordo com [3], tem-se:

$$E[y|x] = E[X\beta + \varepsilon]$$

$$E[y|x] = E[X\beta] + E[\varepsilon]$$

$$E[y|x] = X\beta + 0$$

$$E[y|x] = X\beta$$

$$Var[y|x] = Var[X\beta + \varepsilon]$$

$$Var[y|x] = Var[X\beta] + Var[\varepsilon]$$

$$Var[y|x] = \sigma^2$$

O vetor de dimensão $p+1$, cujos elementos compõem a solução de ajuste da função linear em $\beta_0, \beta_1, \beta_2, ..., \beta_k$, a um conjunto de pontos $(y_1, x_{11}, x_{12}, ..., x_{1k}), ..., (y_n, x_{n1}, x_{n2}, ..., x_{nk})$, pelo método de mínimos quadrados, tem-se a Equação 2.11:

$$\hat{\beta} = (X'X)^{-1}X'y, \tag{2.11}$$

com a condição de que $(X'X)^{-1}$ exista.

2.3.1 Teste de Significância do Modelo

O teste de hipótese tem como base uma estatística de distribuição F, com k e $(n-k-1)$ graus de liberdade, sob H_0. Os k graus de liberdade se devem ao fato de termos k parâmetros das variáveis regressoras. As quantidades para calcular o valor observado da estatística são dispostas na tabela de Análise de Variância (ANAVA). São decorrentes da soma de quadrados total de Y (SQT), divididas em 2 parcelas: soma de quadrados da regressão (SQR_{eg}), e a soma de quadrados dos resíduos (SQE) [3].

$$SQT = SQR_{eg} + SQE \tag{2.12}$$

De acordo com a Equação 2.12, tem-se:

$$\sum_{i=1}^{n}(y_i - \overline{y})^2 = \sum_{i=1}^{n}(\hat{y}_i - \overline{y})^2 + \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

Por definição, os quadrados médios são obtidos dividindo as somas de quadrados pelos respectivos graus de liberdade[11].

Então, para o caso de uma regressão linear múltipla, tem-se:

$$QMR_{eg} = \frac{SQR_{eg}}{k}$$

e

$$QMR_{es} = \frac{SQE}{n-k-1}$$

No modelo de regressão o teste F é aplicado para validar se as variáveis regressoras, contribuem de maneira significativa para explicação da variável resposta. As hipóteses a serem testadas são: $H_0 : \beta_1 = 0, ..., \beta_k = 0$ contra H_1 : pelo menos um dos betas difere de zero. Pose-se dizer que a hipótese H_0 específica que o modelo não é significativo, e em H_1 o modelo é significativo.

A estatística do teste é dada pela Equação 2.13:

$$F = \frac{QMR_{eg}}{QMR_{es}} \sim F_{(k,n-k-1)} \tag{2.13}$$

a hipótese $H_0 : \beta_1 = 0, ..., \beta_k = 0$ é rejeitada quando o $F_{calculado} > F_{tabelado}$ ao nível α de significância, com k e $(n-k-1)$ graus de liberdade. Para $F_{calculado} < F_{tabelado}$, aceita-se a hipótese H_0 ao nível α de significância, atestando que há indícios que não há relação linear entre as variáveis.

A Tabela 2.1 representa a ANAVA:

Tabela 2.1: Tabela ANAVA para o MRLM

FV (Fonte de Variação)	GL (Graus de Liberdade)	SQ (Soma de Quadrados)	QM (Quadrado Médio)	F_0
Regressão	k	SQR_{eg}	$\frac{SQR_{eg}}{k}$	$\frac{QMR_{eg}}{QMR_{es}}$
Erro	$n-k-1$	SQE	$\frac{SQE}{n-k-1}$	-
Total	$n-1$	SQT	-	-

Em um MRLM é interessante testar a significância de cada um dos coeficientes de regressão, ou seja, existe interesse em verificar se a contribuição de uma variável regressora é significativa ou não [3].

Para testar individualmente as variáveis, $X_k, K = 1, 2, 3, ..., p$. Tem-se as seguintes hipóteses: $H_0 : \beta_k = 0$ contra $H_1 : \beta_k \neq 0$.

A estatística do teste é dada pela Equação 2.14:

$$T_{\beta_k} = \frac{\hat{\beta}_k - \beta_k}{S(\hat{\beta}_k)} \sim t_{n-p} \tag{2.14}$$

a hipótese H_0 é rejeitada se $t_{calc} > t_{tab}$ ao nível de significância α, atestando que que a variável foi significativa para explicação do modelo. Para $t_{calc} < t_{tab}$ ao nível de significância α a hipótese H_0 é aceita, conclui-se que a variável não foi significativa para o modelo linear.

2.3.2 Coeficiente de Determinação

O coeficiente de determinação R^2, é interpretado como a proporção da variabilidade dos valores observados de Y, explicada pelo modelo considerado. O valor de R^2 pertence ao intervalo $[0, 1]$, em que quanto mais próximo de 1, melhor o ajuste do modelo considerado [3].

O coeficiente de determinação é calculado de acordo com a Equação 2.15:

$$R^2 = \frac{SQR_{eg}}{SQT} = \frac{\sum_{i=1}^{n}(\hat{y}_i - \overline{y})^2}{\sum_{i=1}^{n}(y_i - \overline{y})^2}, 0 \leq R^2 \leq 1 \tag{2.15}$$

2.3.3 Coeficiente de Determinação Ajustado

O coeficiente de determinação é uma medida descritiva da qualidade do ajustamento obtido. Entretanto, o valor do coeficiente de determinação depende do número de observações da amostra, tendendo a crescer quando n diminui; no limite, para $n = 2$, teríamos sempre $R^2 = 1$, pois dois pontos determinam uma reta e os desvios são, portanto, nulos. Uma forma de contornar essa limitação é usar o coeficiente de determinação ajustado para graus de liberdade, indicado por $R^2_{ajustado}$ [11]. Sabemos que:

$$1 - R^2 = 1 - \frac{SQR_{eg}}{SQT}$$

O coeficiente de determinação ajustado é definido por:

$$1 - R^2_{ajustado} = \frac{\frac{1}{n-p}SQR_{eg}}{\frac{1}{n-1}SQT} = \frac{n-p}{n-1}(1 - R^2) \tag{2.16}$$

ou

$$R^2_{ajustado} = R^2 - \frac{1}{n-p}\left(1 - R^2\right) \tag{2.17}$$

De acordo com a Equações 2.16 e 2.17, excluindo o caso em que $R^2 = 1$, temos $R^2_{ajustado} < R^2$. Note que $R^2_{ajustado}$ pode assumir valores negativos.

2.3.4 MULTICOLINEARIDADE

Outro aspecto importante no ajuste de modelos de regressão linear múltipla é a multicolinearidade. Tem-se como objetivo investigar se há multicolinearidade entre as variáveis regressoras, visto que a forte correlação entre elas pode resultar em efeitos negativos no ajuste do modelo. A multicolinearidade é um problema comum em regressão linear, indicando que existe uma relação de linearidade entre as variáveis independentes, prejudicando a estimação dos coeficientes da regressão [15].

Uma das formas de identificar a presença de multicolinearidade é avaliar o Fator de Inflação da Variância (VIF - *Variance Inflation Factor*). Esse fator mede a associação entre as variáveis regressoras de acordo com o coeficiente de determinação do modelo de regressão, apenas com as variáveis independentes. De acordo com [1] o Fator de Inflação da Variância é definido de acordo com a Equação 2.18 como:

$$VIF_i = \frac{1}{1 - R_i^2} \tag{2.18}$$

em que R_i^2 é o coeficiente de determinação da regressão da variável explicativa X_i sobre as outras variáveis explicativas com $i = 1, 2, ..., k$, sendo k quantidade de variáveis explicativas no modelo [15].

Geralmente, considera-se VIF > 10 como um indicativo de problema de multicolinearidade.

2.4 VARIÁVEL *Dummy*

Ao trabalharmos com os modelos de regressão, podemos ter que utilizar uma variável binária (variável *dummy*), em que é atribuído apenas dois valores distintos, quase sempre representados por 1 ou 0. Podemos caracterizar os níveis de medida de uma variável:

i) Escala nominal, quando temos uma classificação em categorias. Por exemplo: sexo;

ii) Escala ordinal, válida para a ordem dos números. Por exemplo: classe econômica social;

iii) Escalar intervalar, para este caso considera a ordem e também podemos realizar uma comparação numérica, intervalos (diferenças) entre valores. Por exemplo: magnitude dos terremotos medidas na Escala *Richter* e o ano em um determinado país;

iv) Escala cardinal, quando são válidas todas as operações com os valores. Por exemplo: valor monetário.

Para construção de variável *dummy* que representem uma variável nominal A com k categorias, $A_1, A_2, A_3, ..., A_k$, são criadas$(k-1)$ variáveis, $Z_1, Z_2, Z_3, ..., Z_{k-1}$, assumindo valores 0 ou 1 de forma que para $i = 1, 2, 3, ..., k-1$, tenhamos para Z_i:

- 1, se a unidade a ser considerada pertence a categoria A_i, ou;

- 0, se a unidade a ser considerada pertence a categoria $A_j, j \neq i$.

Uma das principais características deste tipo de variável é a sua total arbitrariedade.

2.5 SELEÇÃO DE VARIÁVEIS REGRESSORAS

Para o ajuste de um MRLM é necessário determinarmos quais variáveis regressoras melhor explicam a variável resposta do problema, ou seja, dentro de todas as variáveis possíveis escolher adequadamente aquelas que melhor ajustam o modelo. Utilizando os métodos adequados, a signicância e adequação do modelo deve ser verificada e realizada a análise de resíduos [3].

Existem vários métodos utilizados para a seleção das variáveis regressoras a serem incorporadas ao modelo de regressão. O métodos são os seguintes:

i) todas as regressões possíveis: descrito pelo método de seleção que escolhe entre todos os modelos de regressão possíveis;

ii) método "passo a frente"(*forward*): critério que inicia com o ajuste de um MRLS utilizando as variáreis regressoras, podendo as mesmas serem acrescentadas no modelo;

iii) método "passo atrás"(*backward*): critério que inicia com um ajuste de um MRLM com todas as variáveis, podendo elas serem eliminadas do modelo;

iv) método "passo a passo"(*stepwise*): é uma forma generalizado do método *forward*, que permite de forma alternada, eliminações e inclusões de variáveis no modelo de regressão.

Para o presente trabalho, apresenta-se apenas o método *backward*, que foi utilizado para a realização das análises. Este procedimento caracteriza-se por incorporar, inicialmente, todas as variáveis auxiliares em um modelo de regressão linear múltipla e percorrer etapas, nas quais uma variável por vez pode vir a ser eliminada. Se em uma dada etapa não houver eliminação de alguma variável, o processo é então interrompido e as variáveis restantes definem o modelo final [3].

A estatística do teste é dada pela Equação 2.19:

$$\frac{SQR^c_{eg} - SQR^r_{eg}}{QME^c} \sim F_{(\alpha,1,n-p)} \tag{2.19}$$

em que SQR^c_{eg} e QME^c são calculados de acordo com o modelo de regressão completo e SQR^r_{eg} é calculado por meio do modelo de regressão reduzido. A estatística 2.20 testa a contribuição da variável após a inclusão das demais. A contribuição é significante se o valor da estatística for maior que um quantil especificado da distribuição F com 1 e $(n-p)$ graus de liberdade, sendo p o número de parâmetros do modelo completo. Assim, se o valor da estatística for menor que esse quantil da distribuição F, a contribuição não é considerada significante e o modelo reduzido deverá ser preferido. Se observarmos várias variáveis não significantes, apenas uma variável é eliminada em uma etapa (aquela cuja estatística do teste tiver o menor valor). Quando uma

variável é eliminada, passamos para a nova etapa cujo modelo completo não contém a variável que foi descartada. Se todas as variáveis são significantes, o processo é concluído, e o modelo completo desta etapa é o modelo final [3].

2.6 ANÁLISE DE RESÍDUOS

Segundo [3], os resíduos de um modelo de regressão linear têm uma relação muito forte com a qualidade do ajuste, bem como com a confiabilidade dos testes estatísticos sobre os parâmetros do modelo. Nesse sentido, a análise de resíduos tem uma importância fundamental na verificação da qualidade dos ajustes de modelos. Basicamente, essa análise fornece evidências sobre possíveis violações nas suposições do modelo, tais como a de normalidade, homocedasticidade, e quando for o caso ainda fornece indícios de falta de ajuste do modelo proposto.

O vetor de resíduos é definido por:

$$\varepsilon = Y - X\beta$$

Assim, a esperança dos resíduos definidas respectivamente por:

$$E[\varepsilon] = E[Y - X\beta] = 0$$

e

$$Var[\varepsilon] = \sigma^2[I - X(X'X)^{-1}X']$$

em que I é a matriz identidade.

2.6.1 TESTE DE SHAPIRO-WILK

No modelo de regressão o teste de normalidade é utilizado para determinar se a distribuição dos resíduos é adequada à uma distribuição normal, utiliza-se o teste de Shapiro-Wilk para realizar essa verificação. As hipóteses dos teste são: H_0 : os resíduos seguem normalidade *versus* H_1 : os resíduos não seguem normalidade.

De acordo com [13], a estatística do teste é dada pela Equação 2.20:

$$W = \frac{\left[\sum_{i=1}^{n/2} a_i y_i\right]^2}{\sum_{i=1}^{n}(y_i - \hat{y})^2} \tag{2.20}$$

desde que

$$\left(\sum_{i=1}^{n/2} a_i y_i\right) \leq \sum_{i=1}^{n/2} a_i^2 \sum_{i=1}^{n/2} y_i^2 = \sum_{i=1}^{n/2} y_i^2$$

em que, a_i é o melhor estimador não-viciado do valor esperado das estatísticas de ordem de uma amostra de tamanho n com distribuição normal. Realiza-se o teste de normalidade, rejeita-se H_0 a um nível de significância α se p-valor $< \alpha$. A tabela W indica a porcentagem empírica aproximada dos pontos.

2.6.2 Teste de Durbin-Watson

Para verificar a existência de independência nos resíduos da regressão utilizamos, frequentemente, o teste de Durbin-Watson [11]. As hipóteses para aplicação do teste são: H_0 : os resíduos apresentam independência *versus* H_1 : os resíduos não apresentam independência.

A estatística do teste é dada pela Equação 2.21:

$$d = \frac{\sum_{i=2}^{n} (e_i - e_{i-2})^2}{\sum_{i=2}^{n} e_i^2}, 0 \leq d \leq 4 \tag{2.21}$$

onde os e_i são os desvios da regressão ajustada pelo método de mínimos quadrados.

A distribuição de d depende do tamanho (n) da amostra, do número (p) de parâmetros estimados e também da matriz X. Para tornar mais simples a maneira de efetuar o teste, foram tabelados, para diferentes valores de n e de p, aos níveis de significância de 1% e 5% (unilaterais), intervalos (d_L, d_U) que contêm valor crítico, qualquer que seja a matriz X [11].

A regra do teste consiste em: o valor de d é comparado com d_L E d_U. Se $d < d_L$, o resultado é significativo, aceitando a H_0 em favor de H_1. Se $d > d_U$, o resultado é não significativo, ou seja, rejeita-se H_0. Se $d_L < d < d_U$, o resultado é inconclusivo. O valor de d é comparado com $4 - d_L$ e $4 - d_U$. O resultado é significativo se $d > 4 - d_L$ e não-significativo se $d < 4 - d_U$. Se $4 - d_U < d < 4 - d_L$, o resultado é inconclusivo.

2.6.3 Estatística de Influência

Métodos de diagnóstico são utilizados para que desvios entre as observações e os valores ajustados do modelo sejam analisados e verificados o seu grau de influência sobre a análise. Essa diferença entre o real e o previsto podem surgir por vários motivos, pela função de variância, função de ligação, ausência ou não parâmetro de dispersão, parâmetro de inflação nos zeros, ou ainda pela definição errada da escala da variável ou mesmo porque algumas observações se mostram dependentes ou possuem correlação serial. Discrepâncias pontuais podem ocorrer porque as observações estão nos limites observáveis da variável, erros de digitação ou algum fator não controlado (mas relevante) influenciou a sua obtenção [12].

Note que $H = X(X'X)^{-1}X'$, os elementos da diagonal principal desta matriz, denotados por h_{ii} medem o quão distante a observação y_i está das demais $n-1$ observações do espaço definido pelas variáveis explicativas no modelo e h_{ii} depende apenas do valores das variáveis regressoras relacionadas à matriz X. Portanto o elemento h_{ii} representa uma medida de alavanca da i-ésima

observação [9].

De acordo com [5], a distância de Cook foi proposta como forma de avaliar a influência desses pontos no vetor de parâmetros estimados e consequentemente no vetor de valores preditos pelo modelo.

A equação da distância de Cook é dada pela Equação 2.22:

$$D_i = \frac{h_{ii}}{p(1-h_{ii})} r_i^* \tag{2.22}$$

onde,

$$r_i^* = \frac{y_i - \hat{y}_i}{\sqrt{QMR_{es}(1-h_{ii})}}$$

Observa-se que o valor de D_i será grande quando tivermos a medida de alavancagem h_{ii}, próxima do valor de 1 e quando a medida de discrepância da i-ésima observação dada quando r_i^* assumir um valor grande.

A medida de distância de Cook não será adequada para quando os resíduos padronizados forem grandes e os valores de h_{ii} forem próximos de zero. Outros autores também recomendam como indicador de ponto de alavanca $h_{ii} = \frac{2p}{n}$. Esta regra funciona bem na prática, entretanto, geralmente detecta muitas observações que poderão ser pontos de alavanca ou não. Assim, outras medidas de diagnóstico serão sempre necessárias para confirmar esse primeiro diagnóstico [4].

A distância de Cook é uma medida do quadrado da distância entre a estimativa usual de mínimos quadrados de $\hat{\beta}$ baseada em todas as n observações, e a estimativa $\hat{\beta}_j$ obtida quando o j-ésimo ponto for removido.

Assim a distância de Cook pode ser calculada pela Equação 2.23:

$$D_i = \frac{(\hat{\beta}_j - \hat{\beta})' X' X (\hat{\beta}_j - \hat{\beta})}{p \times QMR_{es}} \tag{2.23}$$

3. Metodologia

3.1 Variáveis Utilizadas

Para o presente trabalho, todos os dados foram coletados no *site* oficial da *National Basketball Association* - NBA. As informações são das temporadas regulares da liga, do período de 2009 a 2019, exceto a temporada de 2011-2012, que foi desconsiderada devido a mesma ter sido completada com 66 jogos totais por equipe, totalizando 9 temporadas. O campeonato é composto por 30 equipes, em que cada uma delas realiza 82 jogos por temporada. O banco de dados possui 270 observações, dais quais cada linha de informações contém as estatísticas para aquela equipe na temporada. Não foram encontradas informações incompletas ou dados faltantes.

A Figura 3.1, ilustra a variável binária CONF, a qual indica em qual conferência pertence a equipe, Leste ou Oeste. Pode-se observar os Estados em cor azul, representados pelas equipes que disputam a conferência Oeste, e os Estados em vermelho representados pelos times que atuam na conferência Leste. A figura ilustra a quais divisões dentro da conferência pertence a equipe, característica que não será discutida no presente trabalho.

Figura 3.1: Mapa das Equipes da NBA

A Tabela 3.1 apresenta a variável resposta número total de vitórias da equipe, e 21 variáveis explicativas para a realização do estudo, em que todas elas são variáveis contínuas.

Tabela 3.1: Variáveis Utilizadas Para as Análises

Estatística	Descrição
W	Quantidade total de vitórias da equipe
PTS	Média de pontos por partida da equipe
FGM	Quantidade média de arremessos convertidos
FGA	Quantidade média de arremessos tentados
FG%	% média de arremessos convertidos
3PA	Quantidade média de arremessos de 3 pontos tentados
3PM	Quantidade média de arremessos de 3 pontos convertidos
3P%	% média de arremessos de 3 pontos convertidos
FTM	Quantidade média de arremessos de lances livres convertidos
FTA	Quantidade média de arremessos de lances livres tentados
FT%	% média de arremessos de lances livres convertidos
OREB	Quantidade média de rebotes ofensivos
DREB	Quantidade média de rebotes defensivos
REB	Quantidade média de rebotes
AST	Quantidade média de assistências
TOV	Quantidade média de bolas perdidas
STL	Quantidade média de bolas roubadas
BLK	Quantidade média de tocos
BLKA	Quantidade média de tocos tomados
PF	Quantidade média de faltas
PFD	Quantidade média de faltas que a equipe tende a receber
PLUS/MINUS	Quantidade média de pontos em relação ao adversário

3.2 Desenvolvimento do Modelo de Regressão

Utilizou-se para o presente trabalho o *software* R [10] para tratamento, manipulação de dados, análises, estimação do modelo de regressão linear múltipla, construção de gráficos diagnósticos e validação dos resultados.

Para a construção do modelo de regressão linear múltipla, considerou-se a variável resposta Y como sendo a quantidade total de vitórias de cada equipe durante a temporada regular. Inicialmente, foram incluídas 22 variáveis regressoras para realizar o ajuste do modelo de regressão, nas quais, 21 eram variáveis contínuas e 1 variável binária.

Foram realizadas estatísticas descritivas da variável resposta e das variáveis explicativas para dimensionamento das características do banco de dados e compreensão das variáveis utilizadas. Em seguida, para a modelagem, os parâmetros do modelo de regressão linear foram estimados de acordo com o método de mínimos quadrados. Utilizou-se o critério de eliminação de variáveis *backward*, o qual considera, para obtenção das variáveis que permanecem no modelo, o teste F descrito na Equação 2.19 para obtenção das variáveis regressoras que permaneceram no modelo.

O teste t-Student a um nível de significância de 5% foi utilizado para verificar a significância das variáveis no modelo final. Também foi utilizado o fator de inflação da variância (VIF), para verificar a multicolinearidade das variáveis.

Além disso, para verificação de que o modelo teve realmente um bom ajuste, foram calculados os valores de R^2 e $R^2_{ajustado}$ do modelo de regressão obtida. Como parte da validação do modelo, foram analisados os gráficos *box plot*, *qq plot* e histograma para uma análise gráfica diagnóstico dos resíduos. Para identificar possíveis pontos de influência, foi feita uma investigação visual por meio de gráfico da distância de Cook. Por fim, foi feito o teste de Shapiro-Wilk e Durbin-Watson para verificar, respectivamente, normalidade e independência dos resíduos da regressão.

4. Resultados

4.1 Estatística Descritiva

A Estatística Descritiva permite resumir, descrever e compreender os dados de uma distribuição usando medidas de tendência central (média, mediana e moda), medidas de dispersão (valores mínimo e máximo, desvio padrão e variância), percentis, quartis e decis. Permite-se com esta ferramenta, concentrar e reduzir as informações. Tais resultados podem ser observados na Tabela 4.1:

Tabela 4.1: Estatística Descritiva da Variável Resposta e Variáveis Regressoras

Variáveis	Mín.	1º Q.	Mediana	Média	3º Q.	Máx.	Var.(σ^2)	Des. Padrão (σ)
W	10	32	42	41	50	73	158.84	12.60
PTS	91.9	98.53	102.45	102.77	106.88	118.10	31.52	5.61
FGM	33.70	37.02	38.30	38.36	39.30	44.00	3.60	1.90
FGA	75.80	81.80	84.00	84.08	86.40	94.00	10.73	3.28
FG%	40.80	44.60	45.50	45.63	46.67	50.30	2.37	1.54
3PM	3.80	6.60	8.40	8.41	10.00	16.10	5.00	2.24
3PA	11.30	18.73	23.25	23.57	27.48	45.40	35.86	6.00
3P%	30.50	34.42	35.50	35.57	36.70	41.60	3.19	1.79
FTM	12.20	16.43	17.50	17.64	18.70	24.10	3.41	1.85
FTA	16.60	21.40	23.10	23.19	24.60	31.10	5.78	2.40
FT%	66.80	74.50	76.20	76.09	78.40	82.80	8.57	2.93
OREB	7.60	9.70	10.60	10.60	11.50	14.60	1.50	1.22
DREB	27.20	30.90	32.40	32.42	33.90	40.40	4.12	2.03
REB	36.90	41.62	43.00	43.03	44.40	49.70	4.41	2.10
AST	18.00	21.00	22.20	22.40	23.57	30.40	4.25	2.06
TOV	11.50	13.60	14.20	14.30	15.00	17.70	1.15	1.07
STL	5.50	7.02	7.60	7.63	8.20	10.00	0.72	0.85
BLK	2.40	4.30	4.80	4.87	5.40	7.60	0.55	0.74
BLKA	3.00	4.40	4.90	4.87	5.40	6.90	0.55	0.74
PF	16.60	19.30	20.40	20.36	21.30	24.80	1.97	1.40
PFD	16.20	19.50	20.40	20.36	21.20	24.30	1.69	1.30
Plus/Minus	-10.500	-3.200	0.200	0.003	3.500	11.600	21.63	4.65

Considerando os dados das temporadas regulares da liga da NBA no período de 2009 a 2019, na Tabela 4.1 tem-se as estatísticas descritivas da variável resposta W (quantidade de vitórias) e das 22 variáveis explicativas que estão sendo analisadas, descrevendo os valores mínimos, 1º quartil, mediana, média, 3º quartil, valores máximos, variância e desvio padrão, respectivamente. A partir desta, podemos avaliar o panorama do banco de dados e suas variáveis para realização das demais análises.

Observando os valores descritos por W, é possível identificar uma variabilidade na variável resposta, com uma variância acima de 150 e um desvio padrão acima de 12, o que condiz com as característica da liga, havendo uma diferença considerável entre as equipes de melhor e pior desempenho durante o campeonato.

Avaliando-se a estatística descritiva da variável PTS, percebe-se que menos de 25% das equipes fazem menos que 98 pontos por partida, ou seja, a contagem centenária nos jogos é algo bastante recorrente durante a temporada, fazendo com que boa parte das equipes privilegiem a parte ofensiva. Considerando as informações coletadas, apenas 15 vezes em 270 observações a equipe com menos de 98 pontos de média obteve uma campanha positiva, que é a quantidade total de vitórias maior que a quantidade total de derrotas.

Na Figura 4.1, observa-se o histograma e o *boxplot* para a variável resposta, nos quais verifica-se uma distribuição próxima da normalidade, não apresentando nenhum valor discrepante (*outlier*).

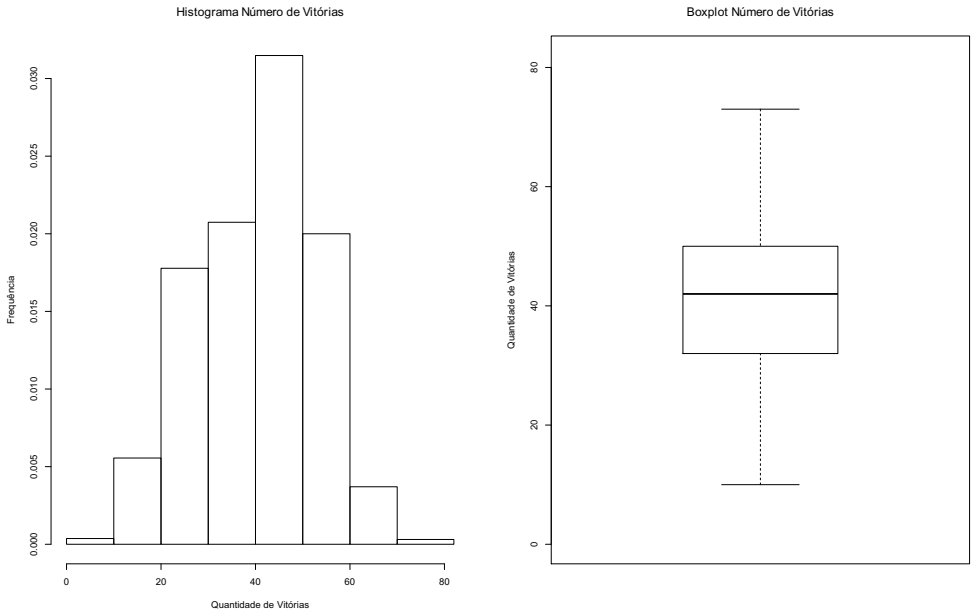


Figura 4.1: Histograma e *Boxplot* da Variável Resposta

Um ponto importante e comum quando se observa a quantidade de vitórias das equipes durante uma temporada da NBA, é a análise e comparação dos resultados de cada conferência. Observando a Figura 4.2, verifica-se que a conferência Oeste possui um valor máximo acima de 70 e um valor mínimo superior a 10 vitórias. Destaca-se ainda que o valor da mediana para

esta conferência fica em torno de 45 vitórias, sendo que o 3º quartil fica acima da frequência observada de 50 vitórias. Pode-se dizer, assim, que mais de 25% das observações do banco de dados desta conferência têm pelo menos 50 vitórias.

Por outro lado, quando se observa a conferência Leste verifica-se algumas diferenças da analise da conferência rival. O gráfico da Figura 4.2 indica que nenhuma equipe ultrapassou a marca de 70 vitórias na temporada, e também que o limite inferior foi de apenas 10 conquistas, o menor valor de todas as observações. Quanto ao valor da mediana para a conferência Leste, observa-se que o valor fica em torno de 41 vitórias e o o 3º quartil do gráfico indica que 25% das observações de vitórias estão iguais ou acima de 48 vitórias.

Por meio da análise descritiva da quantidade de vitórias das equipes de acordo com sua respectiva conferência, pode-se verificar discrepâncias da quantidade total de vitórias entre as equipes que disputam o lado Oeste e as que disputam o lado Leste, conforme pode ser visto na Figura 4.2. Claro que para efeito de qual será a equipe campeã, esta avaliação não tem grande relevância, visto que estamos analisando apenas a temporada regular, ou seja, não inclui a fase *playoffs*, que é a fase final do campeonato.

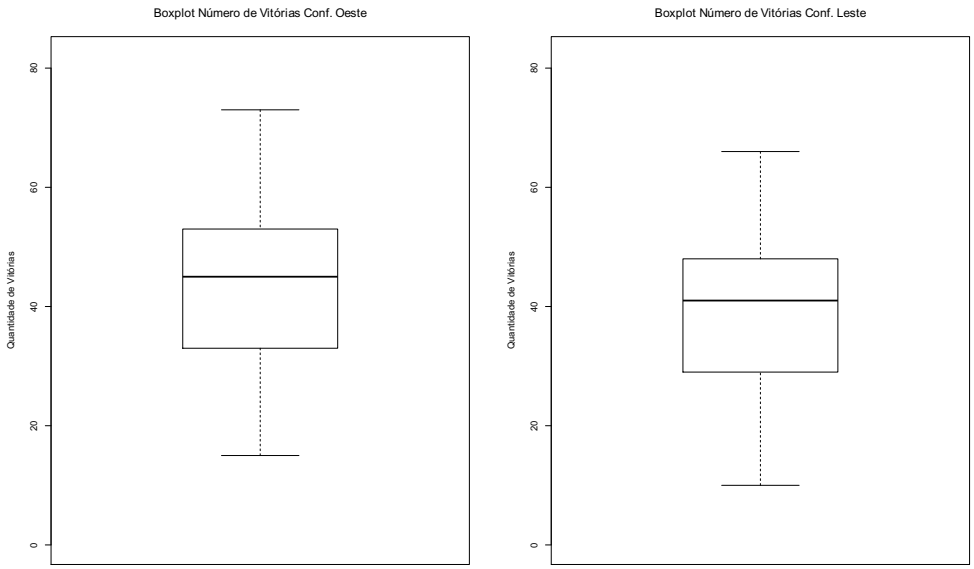


Figura 4.2: *Box Plot* da Quantidade de Vitórias das Equipes de Cada Conferência

Para verificar se houve diferença nas médias de vitórias de duas conferências, realizou-se o teste t para comparação de médias, obtendo-se um p-valor de 0.01228. Concluindo-se assim, que existe diferença significativa entre duas conferências, com relação ao número de vitórias, considerando 5% de significância.

4.2 Estimação do Modelo de Regressão

Uma estatística presente no jogo e que geralmente é interpretada pelos especialistas como sendo de grande importância, é a percentagem de lances livre convertidos. Essa estatística

costuma ser alvo de muita discussão.

A frase "Lance livre ganha jogo", é sempre dita por comentaristas de basquete nos momentos finais de uma partida, quando ocorrem várias faltas e o jogadores vão para a linha de lance livre para tentar converter o arremesso, sem marcação do adversário. É de senso comum que isto acontece realmente, mas não se pode afirmar que este fato acontece de modo linear, em que o aumento da percentagem de lances livres corretos irá realmente aumentar a quantidade de vitórias da equipe.

No exemplo em questão, verifica-se na Figura 4.3 que não há relação evidente entre a porcentagem de lances livres convertidos e a quantidade de vitórias. Sendo bastante subjetiva a análise gráfica, realizou-se o ajuste de um modelo de regressão linear entre essas duas variáveis, para verificar a significância do coeficiente angular desta reta, e consequentemente da variável regressora, porcentagem de lances livres convertidos sobre a variável resposta. Utilizou-se o teste *t-student* a um nível de significância de 0,05 ($\alpha = 5\%$). Conforme a Tabela 4.2:

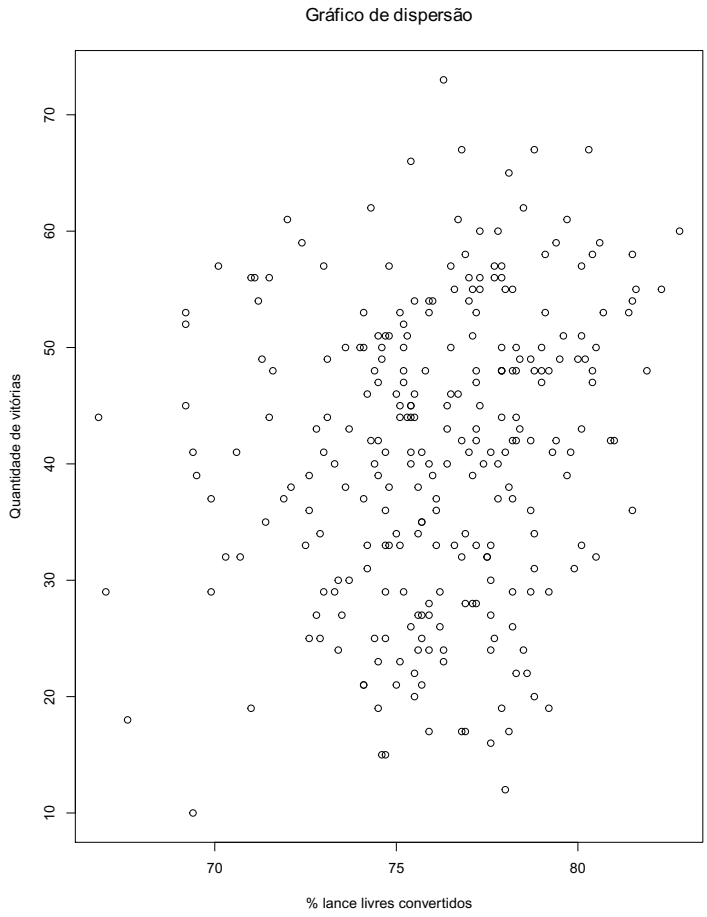


Figura 4.3: Gráfico de Dispersão % Lances Livres por Vitórias

Tabela 4.2: Estimativas Para o Modelo de Regressão Apenas com a Variável FT%

Coeficientes	Estimativa	Erro	Estatística t	p-valor	R^2	$R^2_{ajustado}$
Intercepto	-20.4209	19.6664	-1.038	0.300004	0.03516	0.03516
FT%	0.8072	0.2583	3.125	0.00197		

Observa-se pela Tabela 4.2, que ao fazermos o modelo de regressão linear utilizando apenas a variável FT%, referente a porcentagem média de arremessos de lances livres convertidos, a variável regressora é significativa para explicação da variável resposta. Assim temos que a variável é linearmente relacionada com a variável Y. Porém, é importante destacar que o valor do R^2 e $R^2_{ajustado}$, sendo um valor extremamente baixo para demonstrar que o modelo não apresentou um bom ajuste, sendo necessária a utilização de mais variáveis explicativas para ajuste do modelo de regressão.

Iniciando a análise do modelo completo, com todas as 22 variáveis regressoras, o critério de backward conduziu a um modelo de regressão linear múltipla com apenas 4 variáveis regressoras, sendo elas: FGM, a média de arremessos convertidos na partida; FGA (média de arremessos tentados na partida); TOV (média de bolas perdidas pela equipe na partida); PLUS/MINUS (diferença média de pontos da equipe em relação aos adversários). A significância dos parâmetros estimados estão descritas na Tabela 4.3, verificando-se que de acordo com o critério de *backward*, o modelo mais adequado manteve a variável TOV no modelo, embora esta não seja significativa de acordo com o teste *t-student*. Nesses caso, ocorre geralmente que a exclusão da variável não significativa conduz a um modelo com pior ajuste, sendo mais viável manter o modelo com uma variável não significativa. O coeficiente de determinação deste modelo foi igual a 94,41%, indicando que boa parte da variabilidade da variável resposta pode ser explicada pelo modelo de regressão.

Tabela 4.3: Estimativas Para o Modelo Final de Regressão

Coeficientes	Estimativa	Erro	Estatística t	p-valor	R^2	$R^2_{ajustado}$
Intercepto	51.57800	5.56235	9.273	<2e-16	0.9441	0.9432
FGM	0.37975	0.18844	2.015	0.0449		
FGA	-0.25199	0.10290	-2.449	0.0150		
TOV	-0.27751	0.17976	-1.544	0.1238		
PLUS.MINUS	2.54042	0.05435	46.742	<2e-16		

Para avaliar a presença de multicolinearidade entre as variáveis explicativas, calculou-se o fator de inflação da variância (VIF - *Variance Inflation Factor*), cujos valores são apresentados na tabela 4.4. Como todos os valores foram menores que 10, pode-se concluir que não há multicolinearidade no modelo.

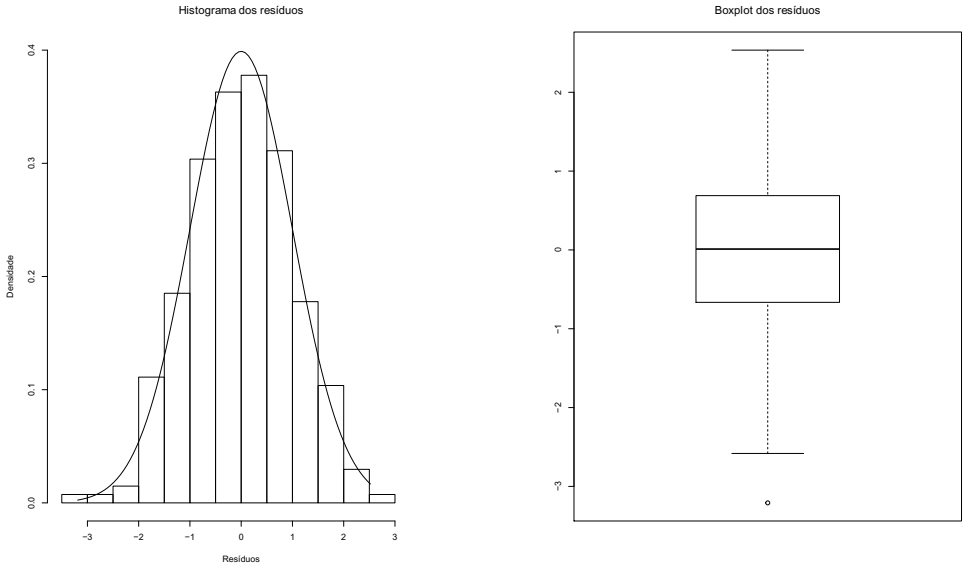
Tabela 4.4: Valores do Fator de Inflação da Variância para as Covariáveis

Variáveis	FGM	FGA	TOV	PLUS/MINUS
VIF	3.733677	3.380905	1.116660	1.863293

4.3 Diagnósticos do Modelo

Para demonstrar o comportamento dos resíduos padronizados do modelo de regressão, analisou-se o histograma e o gráfico de *box plot*, nos quais observa-se apenas um valor dis-

crepante e uma tendência á normalidade, satisfazendo umas das pressuposições do modelo. Tal resultado é representado pela Figura 4.4:



(a) Histograma (b) Box plot

Figura 4.4: Gráficos de Histograma e *Box Plot* para os Resíduos do Modelo

O gráfico *qq-plot* é uma ferramenta muito útil para verificar a adequação da distribuição de frequência dos dados à uma distribuição de probabilidades. Neste caso, verifica-se que os resíduos realmente tendem à distribuição normal, conforme a Figura 4.5:

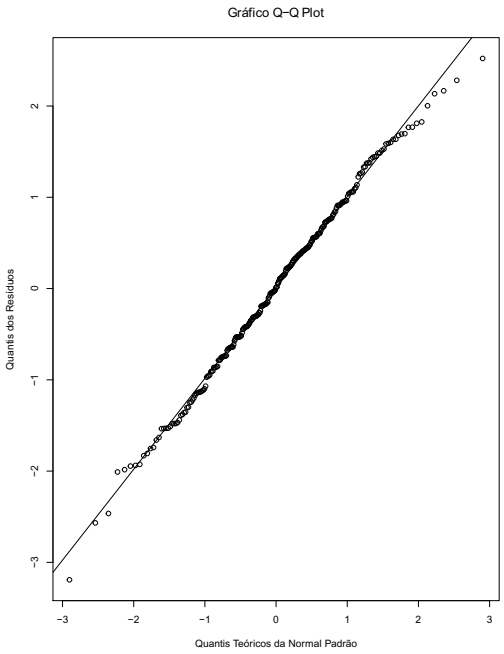


Figura 4.5: Gráfico *QQ-Plot* dos Resíduos

Outra análise importante para verificar o ajuste do modelo de regressão linear, é distância de Cook, que mede a influência da observação sobre sobre o vetor de valores estimados $\hat{Y}$. De

acordo com o gráfico dos valores observados pela distância de Cook, apresentado na Figura 4.6, observa-se 14 pontos discrepantes dentro de um total de 270 observações.

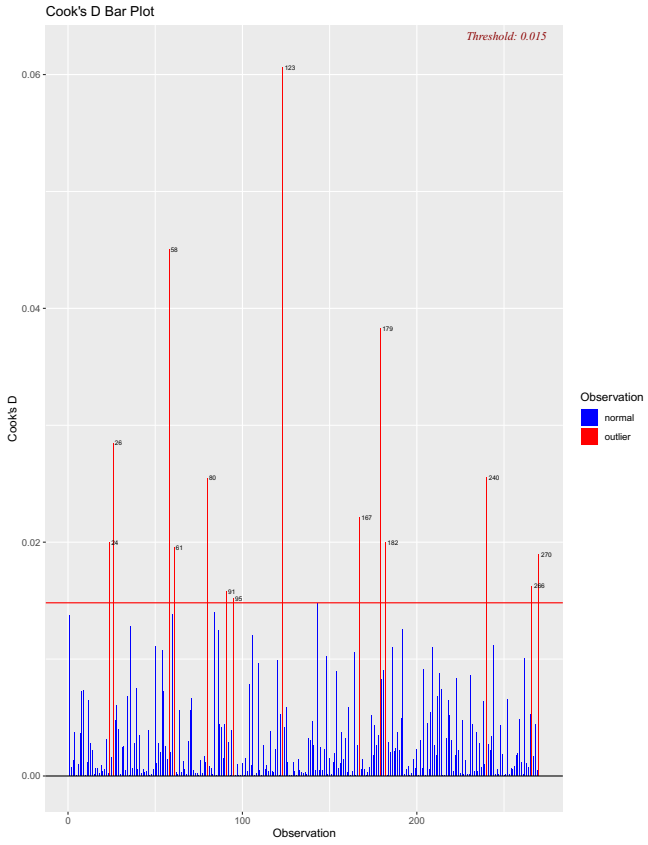


Figura 4.6: Gráfico de Barras dos Valores Observados x Distância de Cook

Ao longo das temporadas da NBA, existem equipes que se destacam positivamente conquistando muitas vitórias, e outras que terminam com poucas vitórias totais, sendo essas observações exemplos de outliers descritos pela distância de Cook. Como exemplo , pode-se citar que dentro das temporadas analisadas, está a campanha da equipe do *Golden State Warriors*, que foi a melhor da história da NBA, totalizando 73 vitórias e apenas 9 derrotas na temporada de 2015-2016. Em contrapartida, também está considerando a temporada do *Philadelphia 76ers*, que teve a segunda pior campanha da história, com apenas 10 vitórias e 72 derrotas na temporada de 2015-2016. Considerando a quantidade de *outliers* observados pelo gráfico, tem-se um baixo percentual (5%) do toral de observações, considerando-se este resultado normal para as temporadas da NBA.

Como forma de verificar a influência dos pontos discrepantes nas estimativas dos parâmetros, ajustou-se um modelo de regressão linear excluindo as estatísticas para estas duas equipes citadas anteriormente, que foram totalmente opostas nos seus desempenhos.As estimativas dos parâmetros para esse novo modelo praticamente não diferiram numericamente das anteriores, havendo apenas poucas alterações na segunda ou terceira casas decimais, levando a crer que estes pontos não se referem a observações influentes.

Uma estratégia comum quando se verifica pontos influentes na aplicação da técnica de análise de regressão, é a análise da retirada destas observações do conjunto de dados, considerando que

estas observações podem influenciar fortemente as estimativas dos parâmetros. Para o modelo de regressão linear apresentado neste trabalho, os pontos discrepantes são os que reforçam a análise fornecendo informações importantes sobre as equipes. Em se tratando de estatísticas de um campeonato de basquete, é normal que se tenha observações de algumas equipes que destoam das demais, portanto não podem ser desconsideradas.

4.4 Testes de Ajuste do Modelo

São apresentados a seguir os resultados das análises dos resíduos do modelo as quais são necessárias para validação do modelo de regressão. Para verificação da homogeneidade das variâncias do modelo, a interpretação será através do gráfico de resíduos. Tais resultados são descritos na Figura 4.7:

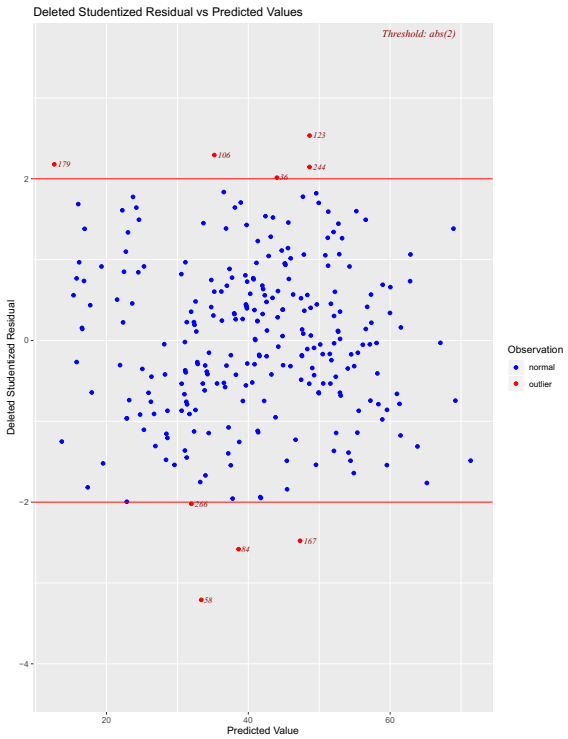


Figura 4.7: Diagrama de Dispersão dos Valores Preditos x Resíduos Padronizados

De acordo com a Figura 4.7, pode-se observar que a distribuição dos resíduos segue constante por todo o modelo, não há uma variação extrema ou concentração de valores de uma determinada área do gráfico, verificando a homogeneidade das variâncias.

Realizando os testes de normalidade e independência dos resíduos, tem-se os resultados de acordo com a Tabela 4.5:

Tabela 4.5: Resultados dos Testes da Análise de Resíduos

Verificar	Teste	Estatística	p-valor
Normalidade	Shapiro-Wilk	0.99607	0.7356
Independência	Durbin-Watson	1.9882	0.4276

De acordo com resultados obtidos na Tabela 4.5, ao nível de significância α igual a 5%, o teste de Shapiro-Wilk não permitiu rejeitar a hipótese H_0, atestando-se a normalidade dos resíduos do modelo de regressão. Observa-se também o resultado do teste de Durbin-Watson, para verificar independência, que ao nível de significância α igual a 5%, não se rejeita a hipótese H_0, atestando-se a independência dos resíduos.

4.5 Interpretação do Modelo

A expressão do modelo de regressão linear múltipla ajustado é apresentada na Equação 4.1:

$$\hat{Y} = 51.578 + 0.37975FGM - 0.25199FGA - 0.27751TOV + 2.54042PLUS/MINUS \quad (4.1)$$

De acordo com as estimativas dos parâmetros do modelo, verificou-se que para a variável FGM referente à quantidade média de arremessos convertidos em uma partida, tem-se uma relação positiva com a variável resposta, ou seja, à medida que a equipe aumenta a quantidade de arremessos convertidos, a quantidade de vitórias também irá acompanhar seu crescimento, mantendo as demais variáveis fixas.

Considerando a variável FGA, referente à quantidade média de arremessos tentados pela equipe durante a partida, observa-se uma relação inversa com a variável resposta, ou seja, mantendo as demais variáveis fixas, à medida que a equipe aumenta a média de arremessos tentados, o número de vitórias tende a diminuir, de acordo com o coeficiente negativo estimado para FGA.

Outra relação inversamente proporcional à variável resposta, é a variável TOV (*turnover*), que se refere à quantidade média de bolas perdidas pela equipe. De acordo com o valor negativo estimado para o coeficiente de TOV, pode-se dizer que à medida que o time diminui a quantidade de ataques perdidos, tem-se um aumento na quantidade de vitórias, mantendo as demais variáveis fixas.

Avaliando a variável PLUS/MINUS, referente à quantidade média de pontos da equipe em relação ao adversário, verifica-se que esta apresentou maior significância no modelo. Considerando o valor positivo da estimativa do parâmetro, tem-se uma relação diretamente proporcional à variável Y, ou seja, à medida que o time aumenta a diferença de pontos em relação ao oponente, a quantidade total de vitória acompanha o seu crescimento, mantendo as demais variáveis fixas.

Observou-se que apenas 4 das 22 variáveis foram mantidas no modelo de regressão ajustado. O modelo também apresentou um coeficiente de determinação alto, mostrando que essas variáveis podem influenciar as equipes a obter mais ou menos sucesso no seu resultado, ou seja, dependendo dos seus valores, podem aumentar ou diminuir o número de vitórias das equipes. Esperava-se que a variável 3PA, que se refere a quantidade média de arremessos de 3 pontos convertidos, fosse significativa, devido as mudanças na forma de como as equipes executam seus ataques, tentando muitos arremessos de 3 pontos.

Notou-se também, que a variável FT% foi significativa quando considerada sozinha no modelo de regressão, e, no entanto, quando considerada em um modelo linear com as demais variáveis, sua significância não foi mantida. Este resultado era esperado parcialmente pelo baixo valor de R^2 no modelo com apenas esta variável regressora, porém, a conclusão não está de acordo com o senso comum, que sempre argumenta que "lance livre ganha jogo"uma vez que foi demonstrado que não é tão determinístico assim.

De acordo com o modelo de regressão que foi ajustado, podem ser obtidas informações das combinações das variáveis, com o objetivo de melhorar a performance da equipe durante a temporada. Uma estratégia sugerida pelo modelo pode ser o treinamento de jogadores que consigam de maneira eficiente converter arremessos, porém, sem fazer um número muito alto de tentativas sem sucesso. Essa estratégia pode ser verificada analisando os resultados de significância da variável FGM e FGA. O treinador também pode realizar ajustes nas jogadas pré combinadas, de maneira que a equipe evite aumentar a quantidade bolas perdidas, TOV, e joguem de forma segura dentro de quadra, evitando perdas de bolas, dando menos oportunidades para equipe adversária realizar contra-ataques. Por fim, é necessário que a equipe seja eficiente nas duas partes da quadra, tanto ataque quanto defesa, levando em conta que o modelo não descreveu significativamente a variável PTS, referente a média de pontos por partida da equipe, e sim a variável PLUS/MINUS, mostrando que não necessariamente a equipe marcando muitos pontos irá sempre obter sucesso, e sim ter um ataque eficiente e uma defesa na mesma proporção.

5. Conclusões

Neste trabalho foram analisadas as principais estatísticas que compõem o jogo de basquete, utilizando o modelo de regressão linear multipla, podendo-se concluir que apenas 4 das variáveis analisadas são significativas para explicar a quantidade de vitórias de uma equipe na NBA. A variável quantidade média de pontos da equipe, como era de se esperar, apresentou relação positiva com a variável resposta, quantidade de vitórias na temporada. Resultado igualmente esperado ocorreu com a variável número médio de arremessos convertidos, ou seja, quanto maior a média de arremessos convertidos, maior o número de vitórias na temporada. No entanto, para a variável quantidade média de arremessos tentados, observou-se uma relação negativa com a variável quantidade de vitórias, ou seja, aumentar a média de arremessos tentados, diminui a quantidade de vitórias, podendo-se concluir que as tentativas não eficientes devem ser evitadas. Algumas variáveis que supostamente podem influenciar no resultado da equipe na temporada não apresentaram significância no modelo, entre elas, a conferência, leste ou oeste, à qual a equipe pertence, e a média de lances livres convertidos. Considerando os critérios de diagnósticos, o ajuste do modelo aos dados foi adequado, explicando 94% da variabilidade da variável número total de vitórias na temporada.

Referências Bibliográficas

[1] Berk, K.: *Tolerance and Condition in Regression Computations.* Journal of the American Statistical Association, 72(360):863–866, 1977.

[2] CBB: *Confederação Brasileira de Basquete*, 2019. <http://www.cbb.com.br/a-cbb/o-basquete/historia-oficial-do-basquete>, acessado em 04 de Maio de 2019.

[3] Charnet, R., L. Freire, C. de, Charnet, E. e Bovino, H.: *Análise de Modelos de Regressão Linear com Aplicações.* Editora da Unicamp, 1ª ed., 1999.

[4] Cordeiro, G. e Neto, E.L. e: *Modelos Paramétricos.* Universidade Federal Rural de Pernambuco, 2006.

[5] D.C.Montgomery, Peck, E. e Vining, G.: *Introduction to Linear Regression Analysis.* John Wiley & Sons, Inc., Hoboken, New Jersey, 5ª ed., 2012.

[6] Lima, W. e Neto, F.d.M.: *O despertar do esporte como negócio*, 2019. <https://www.efdeportes.com/efd181/o-despertar-do-esporte-como-negocio.htm>, acessado em 24 de Abril de 2019.

[7] MATERRS, G.S.: *An Arizona State University Media Enterprise*, 2019. <https://globalsportmatters.com/business/2019/03/07/tv-is-biggest-driver-in-global-sport-league-revenue/>, acessado em 03 de Maio de 2019.

[8] NBA: *The official site of the National Basketball Association*, 2019. <http://www.nba.com/>, acessado em 21 de Março de 2019.

[9] Oliveira, S. de: *Inferência e Análise de Resíduos e de Diagnóstico em Modelos Lineares Generalizados.* Universidade Federal de Juiz de Fora, 2013.

[10] R Core Team: *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria, 2019. https://www.R-project.org/.

[11] R.Hoffmann e Vieira, S.: *Análise de Regressão: uma introdução à econometria.* Hucited, 3ª ed., 1998.

[12] Rossi, A.: *Diagnóstico em Regressão*, 2019. <https://lamfo-unb.github.io/2019/04/13/Diagnostico-em-Regressao/>, acessado em 24 de Novembro de 2019.

[13] Shapiro, S. e Wilk, M.: *An Analysis of Variance Test for Normality*. Biometrika, 52:591–611, 1965.

[14] Sportek, T.: *25 World's Most Popular Sports (Ranked by 13 factors)*, 2018. <https://www.totalsportek.com/most-popular-sports/>, acessado em 26 de Abril de 2019.

[15] Villa, T.E.D.: *Predição do Custo do Milho por Meio de Modelos de Regressão Linear Múltipla*. Universidade Federal de Uberlândia, 2016.

[16] Yang, Y.S.: *Predicting Regular Season of NBA Teams Based on Regression Analysis of Common Basketball Statistics*. University of California, 2015.